DOCUMENT RESUME

ED 445 002 TM 031 538

AUTHOR	Patelis, Thanos; Way, Walter D.; Elliot, Scott
TITLE	Developing a Research Plan for an Online Assessment Program.
PUB DATE	2000-04-00
NOTE	34p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Adaptive Testing; Algebra; Computer Assisted Testing; Higher Education; *Internet; Networks; *Online Systems; Planning; *Research Design; Simulation; Testing Programs
IDENTIFIERS	Large Scale Assessment

ABSTRACT

The purpose of this paper is threefold. First, the description of a research plan for a large-scale computer adaptive testing program migrating from a network to Internet version is presented. The general areas needed in such a research plan include: (1) comparability studies; (2) psychometric analyses; (3) development projects; (4) validity studies; and (5) audits. Second, some data showing the comparability between the network and Internet version of one test (i.e., Elementary Algebra) are presented. Simulations used actual test questions with previously calibrated parameters at known ability groups. The comparability results examine comparability in ability estimates, scale scores, performance levels, course placements, and content specifications. The third aspect of this paper is to suggest the use of the "Association of Test Publisher's Provisional Guidelines" in the development of a research plan. Limitations in the simulation data require additional research. Next steps to replicate the comparability studies are presented. Appendixes contain a discussion of characteristics of the Elementary Algebra test, summaries of item pool parameters, a summary of performance level statements, and an outline of system functionality specifications. (Contains 8 figures and 19 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 445 002

Developing a Research Plan for an Online Assessment Program

Thanos Patelis
The College Board

Walter D. Way
Educational Testing Service

Scott Elliot
Vantage Technologies Knowledge Assessment

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
T. Patelis
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Testing over the Internet: Examples of Types of Uses

Symposium at the Annual Meeting of the
National Council on Measurement in Education
New Orleans, LA
April, 2000

TM031538

BEST COPY AVAILABLE

Table of Contents

Table of Contents ii
Abstract 1
Developing a Research Plan for an Online Assessment Program 2
CBTs on the Internet 2
Research and Development Plan 3
Description of Online Assessment Program 4
Software Functionality over the Internet 6
Comparability 7
Procedures 7
Sample 7
Results 8
Ability Estimates 8
Scale Scores 13
Performance Levels and Placements 16
Content Comparability 18
Conclusions and Next Steps 23
References 25
Appendix A: Characteristics of Elementary Algebra Test 27
Appendix B: Summaries of Pool Item Parameters 28
Appendix C: Summary of Performance Level Statements 29
Appendix D: System Functionality Specifications 30

Abstract

The purpose of this paper is threefold. First, the description of a research plan for a large-scale computer adaptive testing program migrating from a network to Internet version is presented. The general areas one should have in such a research plan include (1) comparability studies, (2) psychometric analyses, (3) development projects, (4) validity studies, and (5) audits. Second, some data showing the comparability between the network and Internet version of one test (i.e., Elementary Algebra) are presented. The comparability results examine comparability in (1) ability estimates, (2) scale scores, (3) performance levels, (4) course placements, and (5) content specifications. The third aspect of this paper is to suggest the use of the Association of Test Publisher's Provisional Guidelines in the development of a research plan. Limitations in the simulation data require additional research. Next steps to replicate the comparability studies are presented.

Developing a Research Plan for an Online Assessment Program

The purpose of this paper is threefold. First, the description of a research plan for a large-scale computer adaptive testing program migrating from a network to Internet version is presented. Second, some data showing the comparability between the network and Internet version of one test (i.e., Elementary Algebra) are presented. The third aspect of this paper is to suggest the use of the Association of Test Publisher's Provisional Guidelines in the development of a research plan. Of course, no research plan is perfect. Therefore, feedback from multiple sources is beneficial.

There are numerous issues facing large-scale programs migrating from paper-and-pencil formats to computer based versions. The test delivery methods may be linear and adaptive, each with overlapping and unique issues that need to be addressed. However, there has been a paucity of discussion involving the issues surrounding the migration of network to Internet versions. The use of the Standards of Educational and Psychological Testing (AERA, APA, & NCME, 1999) (i.e., Joint Standards) and other guidelines are helpful in considering the issues to address. But, the purpose of this paper is to present a sampling of the issues to consider in such a migration, along with some initial data.

CBTs on the Internet

The growth of computer-based tests (CBTs) has extended onto the Internet. Internet access enables radical change in CBTs (Bergstrom & Meehan, 2000). Numerous organizations and companies are putting testing programs on the computer and over the Internet. For example, in June, 1996 the State of California has implemented their employment application and civil service testing process over the Internet (Coffee, Pearce, & Nishimura, 1999). After the first year of implementation 7,533 were tested.

The proliferation of Internet based tests is extending into other areas such as psychology (Harriott, 1997; Buchanan & Smith, 1999) and survey research (Kaplan, 1992). Low-stakes assessments that do not have major security requirements are well suited for the Internet. These programs involve the use of the Internet for administering practice tests, surveys, or non-invasive assessments. In addition to the security considerations of low-stakes assessments on the Internet, the psychometric requirements for such tests can be relaxed, especially when the tests are not adaptive tests.

Technical guidelines for adaptive tests have been suggested (Green, Bock, Humphreys, Linn, and Reckase, 1984). The technical guidelines for computer adaptive tests include examination of dimensionality, measurement error, validity, estimation of item parameters, item pool characteristics, and human

factors. These guidelines were utilized in the development of a research plan for the online assessment.

Using the Internet for assessment has advantages and disadvantages. The advantages include the possibility of removing the restrictions surrounding administration and the use of innovative testing formats (Sampson, 1998). However, Sampson (1998) warned about the disadvantages of testing over the Internet. These disadvantages include (1) problems with unproctored environments, (2) lack of skills in interpreting and using test results by the consumer, and (3) the unqualified administration and interpretation of tests.

Efforts to take advantage of the benefits of adaptive test delivery models have been applied to placement tests (McNabb, 1990; Gordon, 1999). However, there is much diversity in the extent to which placement tests are utilized. As McNabb suggested ten years ago, postsecondary institutions did not fully utilize what placement test offered.

Placement tests have migrated to the Internet. Local (Shermis, Mzumara, Brown, & Lillig, 1997) and national efforts (e.g., ACT and College Board) have implemented computerized placement tests over the Internet. The issues faced with these tests encompass both traditional test standards as represented by the Joint Standards and other guidelines (e.g., Association of Test Publishers' Provisional Guidelines, 2000). Therefore, in addition to technical guidelines, broader issues should be considered.

Research and Development Plan

Recommendations for psychometric research directions provided by Reckase (1989) included suggestions in two areas: (1) Refine current methodology in item functioning across formats and calibrating items and (2) develop methods for modeling person-by-item interactions. These suggestions were both reinforced and expanded by Meijer and Nering (1999).

While many development projects have been suggested by both Reckase (1989) and Meijer and Nering (1999), the issues of a testing program that migrates from a network to an online version involve more immediate concerns. Therefore, in establishing a research plan for an operational computer adaptive test over the Internet, these recommendations were considered, as well as guidelines from the Joint Standards. The research plan for the Internet version of an operational test is composed of the following five components: (1) comparability studies, (2) psychometric properties, (3) development projects, (4) validity studies, and (5) audit.

Comparability studies involve ensuring that the results between the network and Internet versions are similar. The purpose is to show that the scores of each version permit similar placements. Before comparability is shown, it is important to make sure the software including the test delivery model functions at an adequate level. Upon verifying that the software is capable of functioning over the Internet, specific aspects that require evidence of comparability are (1) the ability estimates, (2) the scoring, (3) performance levels, (4) course placements, and (5) content representation.

Psychometric properties include the examinations of dimensionality, reliability, equivalence of test forms, and calibrating pre-test items embedded within the operational assessment program (Green, Bock, Humphreys, Linn, & Reckase, 1984).

Development projects include item development, more efficient IRT models, human-factors research, and introduction of free-response items scored via automated scoring algorithms.

Validity studies involved mostly performing placement validity studies. These are important to undertake because recommended cut-scores may not be applicable to institutions that provide courses with unique requirements and curricula. Therefore, procedures that secure evidence around the course placement rules are both important and recommended.

The audit area represents an effort to maintain the quality of the online assessment program via periodic review. The bases for this audit are the Joint Standards and the Provisional Draft of Guidelines (ATP, 2000). The audit becomes a focal point for the articulation of the research projects that are performed over time, and an opportunity to revise the research plan.

Description of Online Assessment Program

The research plan was developed for a computerized assessment program, ACCUPLACER™. ACCUPLACER™ encompasses assessments administered by computer and paper-and-pencil formats, integrated software for guidance, data management, and reporting. The focus of this research plan, and what concerns this audience, is the computerized test.

Students entering college differ considerably in their skill and proficiency levels in English and mathematics. In addition, 5% of the K-12 public school students (i.e., 2.1 million) are Limited English Proficient (LEP) (NCES, 1997). The college-going rates of all students continue to be on the rise (NCES, 1999). Therefore, there is a need by post-secondary institutions to correctly place students in appropriate courses.

Computerized Placement Tests (CPTs) were developed to determine which course placements are appropriate for college students, and whether remedial instruction is needed. Eight CPTs were developed: Reading Comprehension, Sentence Skills, Arithmetic, Elementary Algebra, College-Level Mathematics, and the Levels of English Proficiency (LOEP). LOEP has three components (1) Reading Skills, (2) Sentence Meaning, and (3) Language Use (College Board, 1997). A description of the characteristics of the Elementary Algebra test is provided in Appendix A. The Elementary Algebra test represents one of the most popular tests among the ACCUPLACER™ tests. Thus, the Elementary Algebra was utilized in this study of comparability between the network and Internet versions.

Detailed description of the development process and psychometric information of the tests are available (see College Board, 1993; 1997). The number of the operational items in the Elementary Algebra pool is 173 items. The number of items administered to an examinee is 12 items.

The CPTs use an adaptive test delivery model using the weighted deviations model (WDM) (Stocking & Swanson, 1993; Swanson & Stocking, 1993). This model treats as constraints the psychometric and content specifications. A list of factors, to be used as constraints, is established (e.g., content and sub-content area, demographic references, etc.). Each factor is weighted according to the importance assigned to meeting these specifications. Items are selected in a way that minimizes the sum of the weighted deviations across all these factors.

The items have been calibrated using the three-parameter logistic IRT model and LOGIST (Wingersky, Patrick, & Lord, 1995). The distributions of the item parameters for the Elementary Algebra pool are shown in Appendix B. Student reports provide the Total Right Score, Range, and Percentile Rank. Based on the number of items the examinee answered correctly from the total number presented, the Total Right Score was calculated using a formula representing what the score is if the student had taken 120 items.

There are three ways in which the post-secondary institution may derive placement information from examinee performance on the CPTs. The first manner is to use the proficiency statements provided for each test (College Board, 1997). See Appendix C for a summary of these proficiency statements for the Elementary Algebra test used in this study. The second manner is to utilize the Percentile Rank information. The third manner is to perform placement validity studies locally at the institution.

The content specifications for the Elementary Algebra test are shown in Table 1. As it can be seen in Table 1, the majority of items in the pool represent Algebraic Expressions and Equations, Inequalities & Word Problems (i.e., 92%).

Table 1
Distribution of Items by the Elementary Algebra Test Specifications

Content	Frequency	Percent
Signed Numbers and Rationals	14	8.09
Algebraic Expressions	82	47.40
Equations, Inequalities & Word Problems	77	44.51
Total	173	

Software Functionality over the Internet

Before any issue of comparability was examined, the functionality of this assessment program over the Internet needed to be examined. The technical differences and conditions of the Internet over a local area network may affect the performance of the assessment system. The type of performance that was considered at this point was the responsiveness of the system to multiple users over a large geographic area with multiple configurations. As a result, a set of criteria, as shown in Appendix D, were developed to examine the functionality of the Internet version of this assessment program.

The critical element of this assessment program both psychometrically and technologically was the test delivery model (i.e., CAT algorithm). Unlike the network version, multiple institutions would be tapping into the CAT algorithm for the selection of items for multiple examinees. Therefore, an accelerated set of requirements emerges that the CAT must handle.

As a result, some modification of the CAT algorithm to process such a large volume of interactions was necessary. The criteria in Appendix D were used to make the necessary modifications in the software, so that the assessment program can function at optimum levels.

In addition, the Internet version provided the opportunity to maintain actual exposure rates. Therefore, instead of using a conditional exposure estimate, the Internet version of the system provided a "live" counter for exposure rates. The effect of this type of exposure rate has not been investigated, and is not part of this study. However, future research might explore the implications of such a dynamic method of exposure control.

Comparability

Once the assessment program was moved from a network to Internet format, the immediate issue of compatibility was raised. In developing the research plan to address comparability the following two general aspects were considered accuracy of scores, and content coverage. Data simulations using the Elementary Algebra test are presented.

The accuracy of scores was examined first. The aspects that were considered in this study were comparability in (1) ability estimates, (2) scale scores (i.e., scoring), (3) performance levels, and (4) course placements.

In terms of content coverage, this study examined the degree to which the (1) the intended content specifications were represented by each version, (2) the comparability of content representation by each version, and (3) the comparability of content representation within the test session by each version.

Procedures

Sample

The sample was comprised of simulations using actual test questions with previously calibrated parameters at known ability groups. Two sets of simulations were performed. One set was performed to examine the comparability of the overall ability estimate. The second set of simulations was performed to examine the comparability of covering the content specifications between versions. For this study, the items from the Elementary Algebra test were used.

For the first set of simulations, one hundred simulees were generated at thirteen ability groups. The ability groups ranged from –3.0 to 3.0 at 0.5 $\underline{\theta}$ unit increments. Items using the CAT algorithm with right/wrong determined based on the simulee's ability level. For each item presented, the probability of an examinee at the specified ability level getting the answer right or wrong was determined from the item characteristic curve (ICC), yielding a number between zero and one. A random number ranging from zero to one was then generated. If the random number was equal to or less than the probability of getting correct, the simulee was recorded as getting the item correct. This procedure was repeated for each item until the fixed length test sequence was completed. A maximum likelihood estimate (utilizing the Newton-Raphson procedure) of

the simulee's ability was determined. The ability estimates were constrained to be between –5 and 5 $\underline{\theta}$ units.

For the second set of simulations involving the Elementary Algebra test, similar methods were used. However, there were some differences in the simulation procedure for the examination of the content specifications[1]. First, the ability groups ranges from –2.0 to 2.0 at 0.5 $\underline{\theta}$ unit increments. In addition, the number of simulees differed between versions. Table 2 summarized the number of simulees for each ability group and version of the test involved in the content specification aspect of this study.

Table 2
Summary of the Number of Simulees by Ability Group and Version
For Content Specification Aspect of Study
Elementary Algebra Test

Ability Group	Network	Internet
-2.0	1,371	100
-1.5	147	100
-1.0	1,210	100
-0.5	137	100
0.0	945	100
0.5	115	100
1.0	501	100
1.5	25	100
2.0	133	100

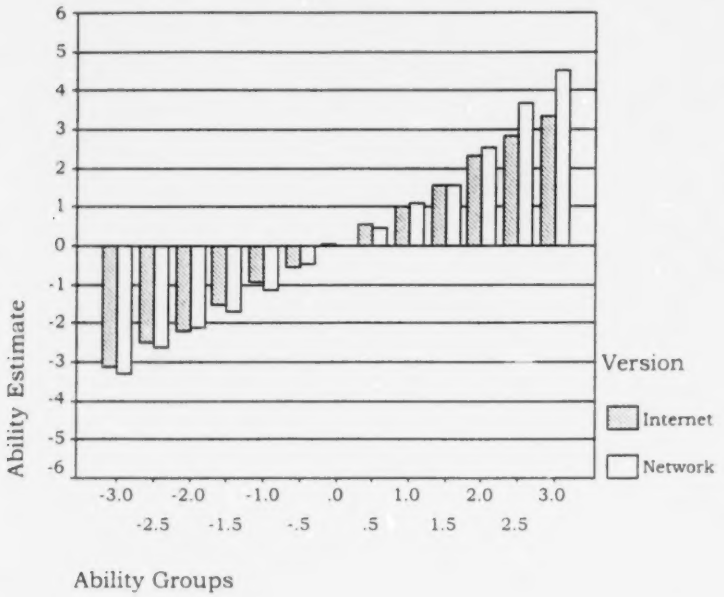
Results

Ability Estimates

A comparison of the mean ability estimates at each of the thirteen ability groups by version for the Elementary Algebra is shown in Figure 1. As can be seen in Figure 1, the ability estimates are very similar between the network and Internet version of the Elementary Algebra test. There are slight differences at the two highest ability groups (i.e., 2.5 and 3.0 $\underline{\theta}$ levels).

[1] The difference in the procedures was based on the different sources used for simulation data for the content specification aspect of this study. No theoretical rationale was involved.

Figure 1
Comparison of the Mean Ability Estimates at each Ability Group
By Network and Internet Version for Elementary Algebra



More detailed descriptive statistics of the ability estimates derived from the first set of simulations are shown in Table 3.

Table 3
Descriptive Statistics of Ability Estimates
By Version and Ability Group
Elementary Algebra

Version	Actual	Mean	N	SD	Minimum	Maximum
Online	-3	-3.13	100	0.98	-5.00	-0.86
	-2.5	-2.51	100	0.89	-5.00	-0.92
	-2	-2.18	100	0.86	-5.00	-0.84
	-1.5	-1.52	100	0.54	-3.37	0.13
	-1	-0.94	100	0.39	-2.62	-0.10
	-0.5	-0.54	100	0.38	-2.19	0.47
	0	0.04	100	0.35	-0.75	1.18
	0.5	0.56	100	0.29	-0.19	1.25
	1	1.02	100	0.30	0.38	1.82
	1.5	1.56	100	0.58	0.44	5.00
	2	2.32	100	0.78	1.20	5.00
	2.5	2.82	100	0.84	1.74	5.00
	3	3.34	100	0.90	1.87	5.00
Network	-3	-3.28	100	1.22	-5.00	-0.61
	-2.5	-2.63	100	1.01	-5.00	-0.76
	-2	-2.11	100	1.07	-5.00	-0.47
	-1.5	-1.67	100	0.75	-5.00	-0.42
	-1	-1.16	100	0.60	-5.00	-0.17
	-0.5	-0.45	100	0.47	-1.37	1.22
	0	0.02	100	0.35	-0.94	1.03
	0.5	0.46	100	0.29	-0.14	1.10
	1	1.09	100	0.34	0.08	1.87
	1.5	1.57	100	0.48	0.91	5.00
	2	2.54	100	1.22	1.39	5.00
	2.5	3.69	100	1.44	1.53	5.00
	3	4.52	100	1.06	1.87	5.00

The results of a two-way (version and ability group) ANOVA found a significant two-way interaction between the version of the test and ability group on ability estimates ($F(12, 2574) = 14.32$, $p < .001$, ES = 0.06). The source table for this ANOVA is shown in Table 4. Included in the source table is the effect size (ES)

of each effect[2]. As can be seen the ES for version and the interaction term (V*AG) were very small suggesting no practical significance for these effects.

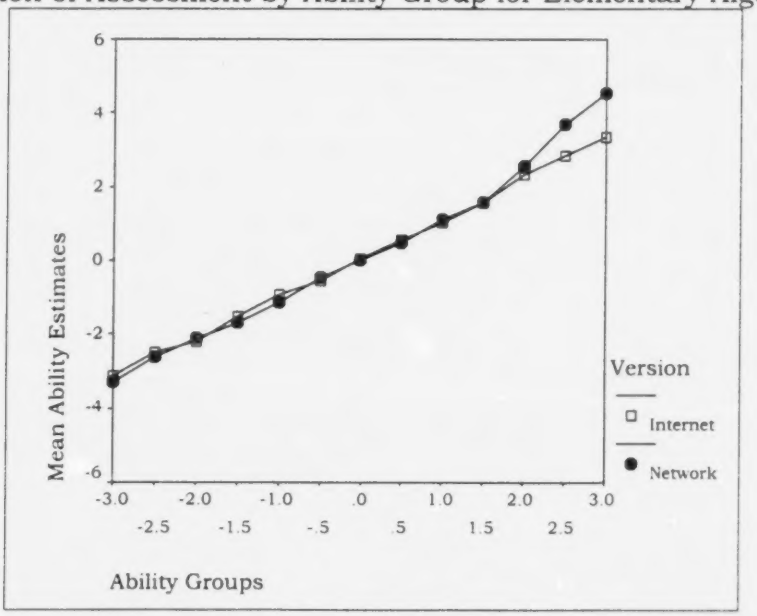
Table 4
Source Table of the Two-Way ANOVA Results Examining the Effects Of Version and Ability Group on Ability Estimates

Source	SS	Df	MS	F	ES
Version (V)	11.73	1	11.73	19.24*	.01
Ability Group (AG)	12,190.66	12	1015.89	1,666.05*	.89
V*AG	104.79	12	8.732	14.32*	.06
Error	1,569.52	2,574	0.61		
Total	13,921.66	2,600			

*p < .01

In order to understand the nature of this interaction, the profile plot of the marginal mean ability estimates were examined. As shown in Figure 2, there is an ordinal interaction at the highest two ability groups of 2.5 and 3.0.

Figure 2
Profile Plot of the Marginal Ability Estimates of Version of Assessment by Ability Group for Elementary Algebra



The differences between the network and Internet versions at these ability groups, however, are practically speaking quite small. When we apply the

[2] The effect size was represented using eta-squared.

conversion table translating the ability estimates to the scaled scores associated with the Elementary Algebra test, the scale scores show no difference between the two versions. The translation of the ability estimates to the scale scores for the two versions is shown in Table 5.

Table 5
Scale Scores for the Mean Ability Estimates from
The 2.5 and 3.0 Ability Groups
Elementary Algebra

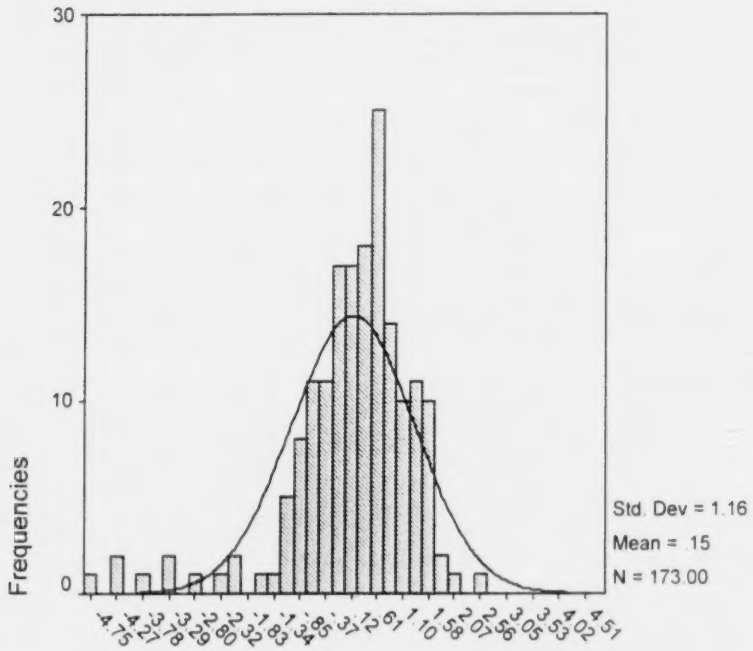
Version	Ability Group	Mean Ability Estimate	Scale Score
Network	2.5	3.69	120.00
	3.0	4.52	120.00
Internet	2.5	2.82	120.00
	3.0	3.34	120.00

The Elementary Algebra test designed to be effective at low to moderate levels of performance. The mean of the maximum information of each item in the Elementary Algebra pool (n = 173) was 0.15. The maximum information (θ_{max}) for each item was calculated by the following equation (Birnbaum, 1968):

$$\theta_{max} = b_i + \frac{1}{Da_i} \ln .5\left(1 + \sqrt{1 + 8c_i}\right)$$

As seen in Figure 3, the distribution of θ_{max} values for the Elementary Algebra test is a negatively skewed distribution with a mean θ_{max} value of 0.15.

Figure 3
Distribution of $\underline{\theta}_{max}$ Values for the Elementary Algebra Test

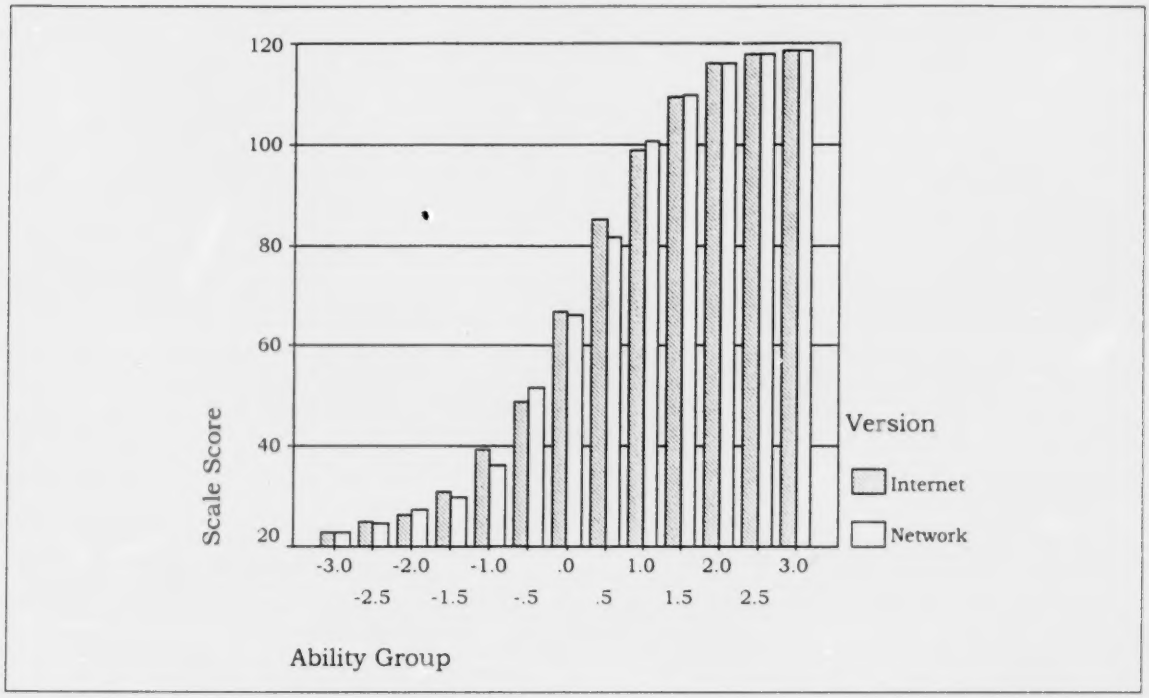


In order to remove the influence of the items with extreme values in our examination of the frequency distribution of the $\underline{\theta}_{max}$ values, the middle fifty percentile of the distribution of $\underline{\theta}_{max}$ values was examined. The middle fifty percentile provided a mean of 0.30. Thus, in examining both distributions, the items were better suited for examinees with slightly better than average ability (i.e., $\underline{\theta} > 0$). It follows that estimated abilities for middle ability groups would be more precise than estimated abilities for candidates in the extreme ability groups. This result is illustrated in the standard deviations of estimated abilities shown in Table 3.

Scale Scores

A second concern regarding the comparability between the network and Internet versions is whether the scoring is similar. Using the conversion table established for converting ability estimates to the scale score of this assessment (20 to 120), the mean scale scores are compared between the two versions in Figure 4.

Figure 4
Comparison of the Mean Scale Scores at each Ability Group
By Network and Internet Version
Elementary Algebra



As can be seen in Figure 4, based on the scale scores, the largest differences between the Internet and network versions can be seen in the –1.0, -0.5, and 0.5 ability groups.

The descriptive statistics shown in Table 6 provide the information needed to make detailed comparisons. In addition to the mean differences between the two versions at the indicated ability groups, the variability was largest in the middle ability groups for each version.

Table 6
Descriptive Statistics of the Scale Scores
By Version and Ability Group
Elementary Algebra

Version	Ability Group	Mean	N	SD	Minimum	Maximum
Internet	-3	22.66	100	2.92	20.03	40.03
	-2.5	24.78	100	4.13	20.03	38.03
	-2	26.33	100	4.60	20.03	40.03
	-1.5	30.83	100	7.27	21.03	70.03
	-1	39.42	100	7.52	23.03	62.03
	-0.5	48.94	100	9.67	24.03	82.03
	0	66.98	100	11.87	42.03	104.03
	0.5	85.00	100	9.87	59.03	106.03
	1	99.00	100	7.62	79.03	114.03
	1.5	109.54	100	5.70	81.03	119.03
	2	116.20	100	3.02	105.03	119.03
	2.5	117.98	100	1.39	114.03	119.03
	3	118.73	100	0.64	115.03	119.03
Network	-3	22.98	100	4.07	20.03	46.03
	-2.5	24.50	100	4.04	20.03	42.03
	-2	27.55	100	5.67	20.03	50.03
	-1.5	29.86	100	5.85	20.03	51.03
	-1	36.21	100	7.70	20.03	60.03
	-0.5	51.60	100	14.00	31.03	105.03
	0	66.17	100	11.64	38.03	100.03
	0.5	81.58	100	9.96	60.03	102.03
	1	100.52	100	8.81	68.03	115.03
	1.5	109.90	100	4.82	97.03	119.03
	2	116.08	100	2.51	108.03	119.03
	2.5	118.01	100	1.52	111.03	119.03
	3	118.72	100	0.84	115.03	119.03

The results of a two-way (version and ability group) ANOVA found a significant two-way interaction between the version of the test and ability group on the scale scores ($F(12, 2574) = 2.88$, $p = .001$, ES = 0.01)[3]. The source table for this ANOVA is shown in Table 7. Included in the source table is the effect size (ES) of each effect[4]. As can be seen the ES for version was zero and the interaction

[3] Because the relationship between the ability estimates and scale scores was non-linear, a two-ANOVA was performed on the scale scores to examine any potential significant effects.
[4] The effect size was represented using eta-squared.

term (V*AG) were very small suggesting no practical significance for these effects. In addition, when the profile plots were examined for the version of the test and ability groups, no discernable difference existed on marginal means of the scale scores.

Table 7
Source Table of the Two-Way ANOVA Results Examining the Effects
Of Version and Ability Group on he Scale Scores

Source	SS	Df	MS	F	ES
Version (V)	28.25	1	28.25	0.57	.00
Ability Group (AG)	3,718,330.30	12	309,860.86	6,248.11*	.97
V*AG	1,711.64	12	142.64	2.88*	.01
Error	127,651.65	2,574	49.59		
Total	16,449,081.08	2,600			

*p < .01

In order to obtain a sense of the sample data to a national reference (College Board, 1993), Table 8 provides the descriptive statistics of the national reference data and the samples representing the Network and Internet versions.

Table 8
Comparison of the Network and Internet Versions
To a National Reference
Elementary Algebra

	National Reference	Network	Internet
M	48.29	69.51	69.72
SD	26.21	38.58	38.38
n	67,263	1,300	1,300

As can be seen in Table 8, the mean values of the national reference were smaller than either sample that represented each version. In addition, the variability, as represented by the SD, was smaller than the SD of each other sample. These statistics suggested that the typical examinee performs at lower levels on this test, and a restriction of range exists.

Performance Levels and Placements

Using the suggested performance levels for Elementary Algebra from Appendix C, the scale scores were translated into performance levels. A comparison

between the performance levels obtained from each version is shown in Table 9.

Table 9
Comparison of the Distribution of Simulees at each Performance Level
By Internet and Network Versions
Elementary Algebra Test

	VERSION				Total
	Internet		Network		
	N	%		%	N
Below Minimal Pre-Algebra Skills	196	15.1%	200	15.4%	396
Minimal Pre-Algebra Skills	399	30.7%	388	29.8%	787
Minimal Elementary Algebra Skills	95	7.3%	116	8.9%	211
Sufficient Elementary Algebra Skills	220	16.9%	206	15.8%	426
Substantial Elementary Algebra Skills	390	30.0%	390	30.0%	780
	1300		1300		2600

As indicated earlier these are suggested performance levels providing one mechanism to interpret the performance of examinees. There were no statistical significant differences in the distribution of performance between the two versions ($\chi^2 = 2.74$, $df = 4$, $p > .05$). Therefore, similar evaluations of performance may be made based on the two versions of this test.

Prescriptions for course placement are not made, because each institution may have different course requirements and curricula. However, the College Board makes suggestions for course placements, and institutions utilize these suggestions. The recommended placement rules based on the Elementary Algebra test are as follows: Examinees with scores between 31 and 56 may be placed into an "elementary algebra I" course; examinees with scores between 57 and 75 may be placed in an "elementary algebra II" course; and examinees with scores between 75 and 107 may be placed in an "intermediate algebra" course. Examinees scoring below 31 are administered the Arithmetic test. Examinees that score above 107 are administered the College-Level Mathematics test. Evidence in support of these recommended cut-score intervals for course placements have been provided recently (Sireci, Patelis, Rizavi, Dillingham, Rodrigues, 2000).

Table 10 shows the distribution of recommended course placements based on the Internet and network versions. There are no statistically significant differences in the course recommendations between these two versions ($\chi^2 = 3.01$, $df = 4$, $p > .05$).

Table 10
Comparison of the Distribution of Simulees
Recommended for Course Placement by Internet and Network Versions
Elementary Algebra Test

Placement Recommendation	VERSION				Total
	Internet		Network		
	N	%		%	N
Administer Arithmetic Test	345	26.5%	330	25.4%	396
Elementary Algebra I	250	19.2%	258	19.8%	787
Elementary Algebra II	95	7.3%	116	8.9%	211
Intermediate Algebra	220	16.9%	206	15.8%	426
Administer College-Level Math. Test	390	30.0%	390	30.0%	780
	1300		1300		2600

Content Comparability

In this section, the degree to which the content specifications were represented by each version of the Elementary Algebra test is presented. Comparisons are made between the Internet and network version, and what is expected by design.

As indicated in Table 2, there are three general content specifications: (A) Signed Numbers and Rationals, (B) Algebraic Representations, and (C) Equations, Inequalities, and Word Problems. The twelve items for this test are selected in varying proportions across these three areas. The distribution of the items selected varies depending on the proficiency of the examinee (See Appendix A for more details).

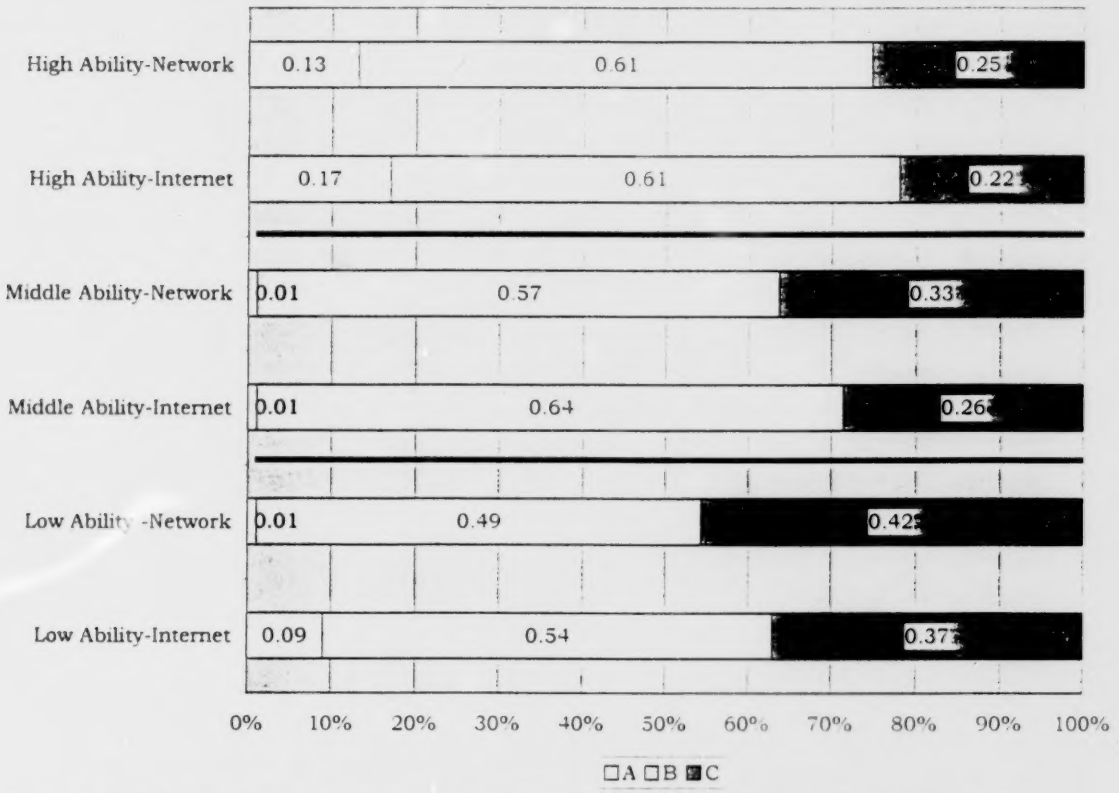
Using the descriptions from Appendix A, the expected distributions of items across the three areas are suggested in Table 11.

Table 11
Suggested Distributions of Items Across the
Three Areas of the Elementary Algebra Test by Low and High Ability

Area	Low Ability	High Ability
A – Signed Numbers & Rationals	25%	8%
B – Algebraic Expressions	58%	42%
C – Equations, Inequalities, & Word Problems	17%	50%

As suggested in Table 11, the expected distribution of the items across areas differed depending on the ability of the examinee. Therefore, more of the difficult items in area C (Equations, Inequalities, & Word Problems) were suppose to be administered to examinees of higher ability.

Figure 5
Comparison of the Distribution of Items by Content Area and Ability Group For Each Version of the Elementary Algebra Test



The nine ability levels from the simulations were condensed in three ability levels (i.e., low, middle, and high ability). This was done by combined the lower three ability levels (i.e., -2.0, -1.5, and -1.0) into the low ability group. Then, the next three ability groups (-0.5, 0.0, and 0.5) were combined into the middle ability group. Finally, the highest three ability groups (1.0, 1.5, and 2.0) were combined into the high ability group. Using these groupings, the mean proportion of items presented at each category for each version was calculated and graphed in Figure 5.

As can be seen in Figure 5, there were some differences between the proportions of items by area at the low and high ability levels and those expected (Table 11). For example, 25% of the items from content area A (Signed

The College Board

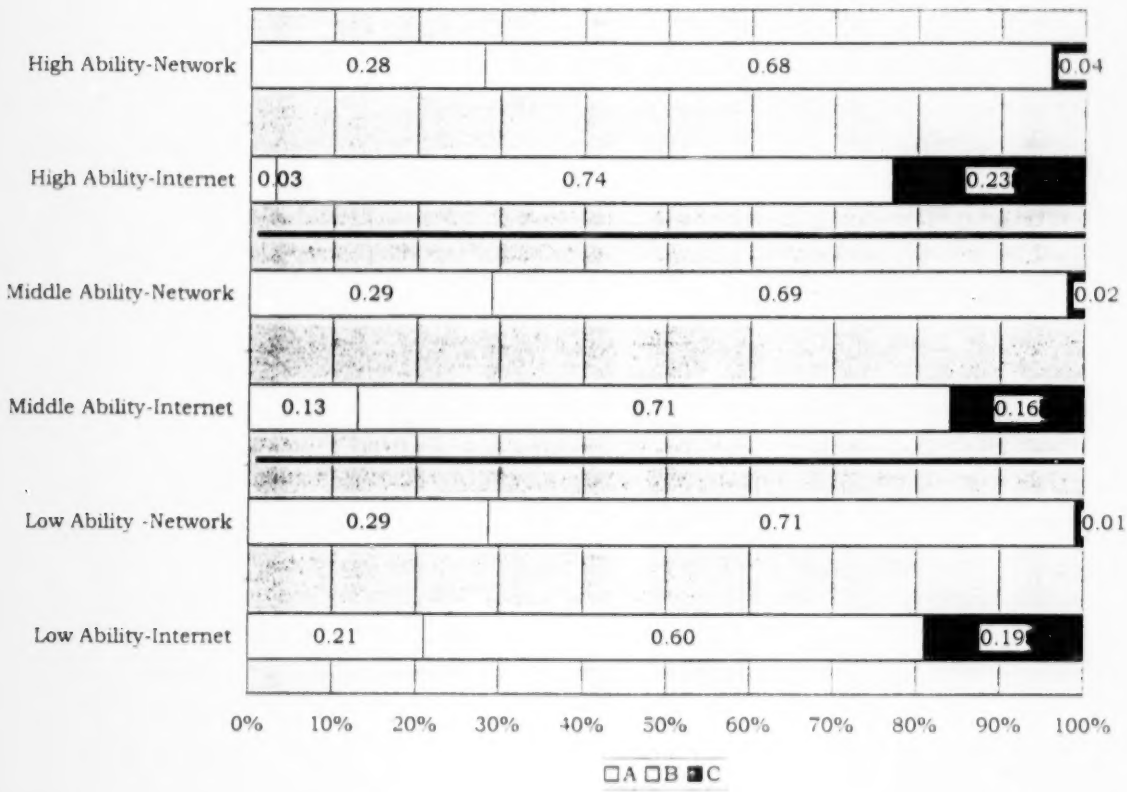
BEST COPY AVAILABLE

Numbers & Rationals) were expected to be presented to low ability examinees. However, both the network and Internet versions presented lower proportions of items from area A to low ability examinees (i.e., 17% and 13% for the Internet and network versions, respectively).

As seen in Figure 5, there are slight differences between the network and Internet versions in terms of the proportions of items selected by area and ability group. However, the differences between the versions were smaller than the differences between each version and the expected distribution.

The final comparisons between the network and Internet versions involved comparing the proportions of items selected by area with parts of the test. To do this, the test was divided into three parts, beginning, middle, and end. Each part consisted of one-third of the items (i.e., four). Figures 6 through 8 show the comparison of the proportions of items by area and version for the beginning, middle, and end of the Elementary Algebra test, respectively.

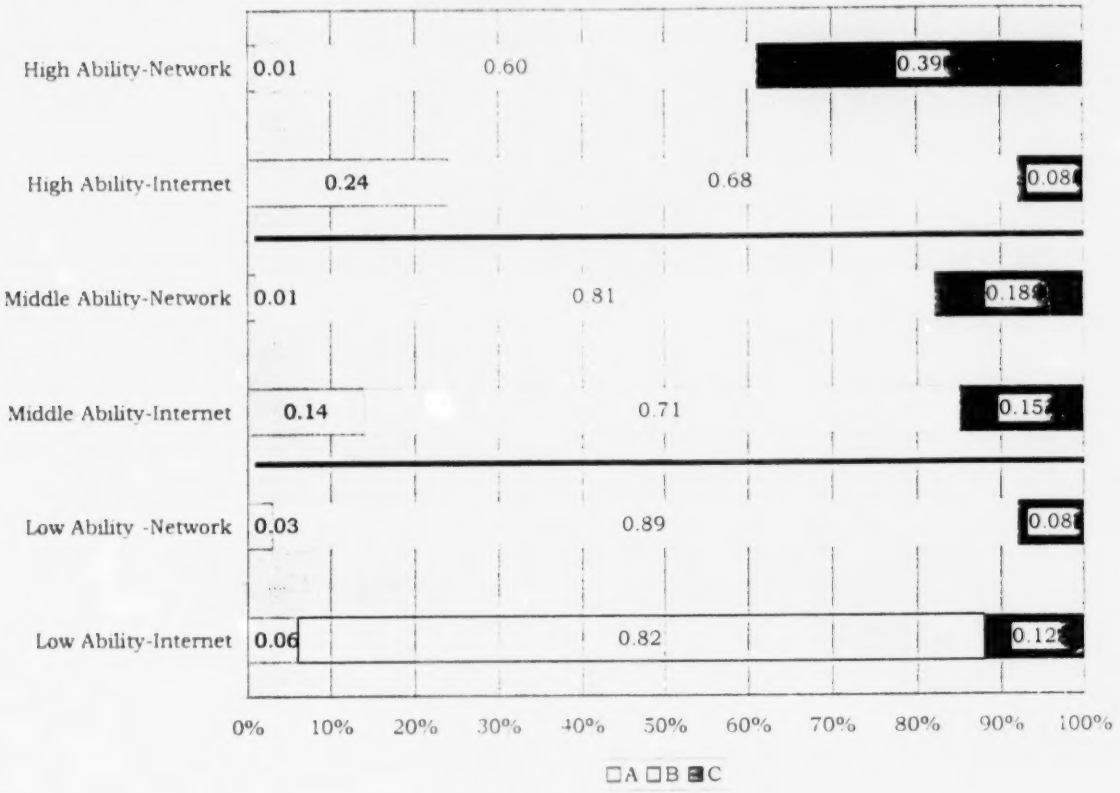
Figure 6
Comparison of the Distribution of Items by Content Area and Ability Group
For Each Version of the Elementary Algebra Test
The Beginning (First Four Questions)



The College Board

BEST COPY AVAILABLE

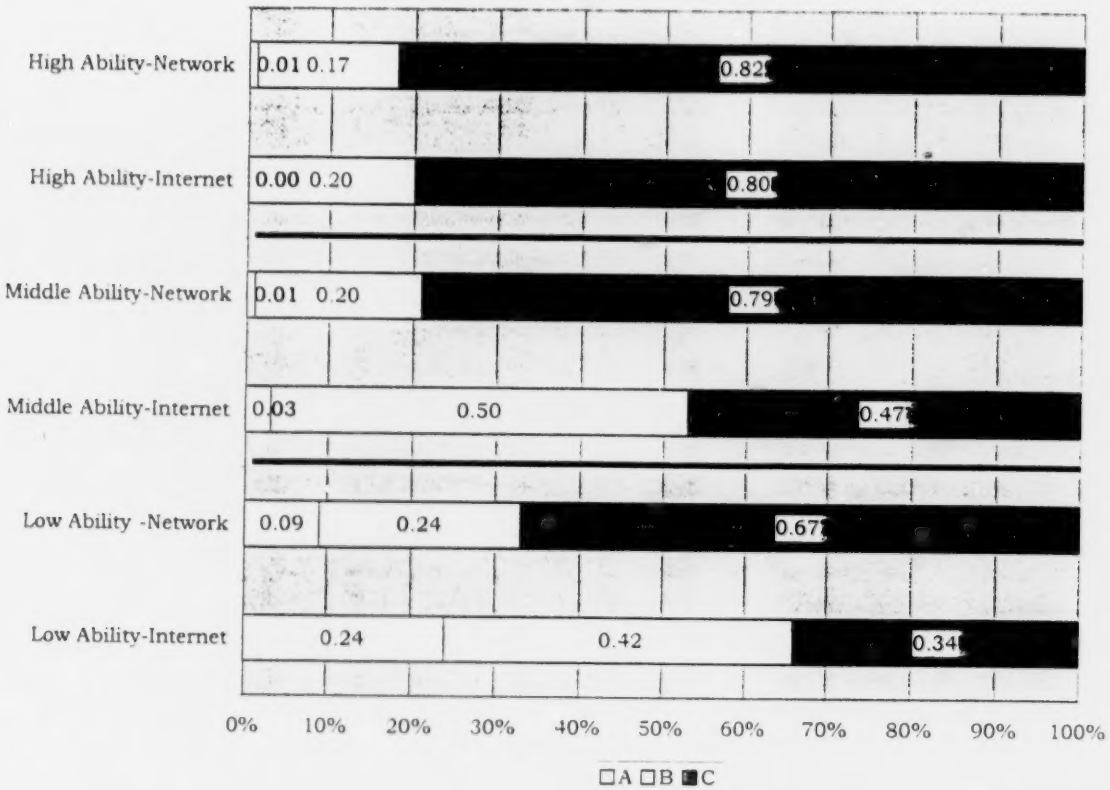
Figure 7
Comparison of the Distribution of Items by Content Area and Ability Group
For Each Version of the Elementary Algebra Test
The Middle (Questions #5 - #8)



The College Board

BEST COPY AVAILABLE 24

Figure 8
Comparison of the Distribution of Items by Content Area and Ability Group
For Each Version of the Elementary Algebra Test
The End (Questions #9 - #12)



As can be seen, there are some differences between the proportion of items across versions of the Elementary Algebra test. Even though the overall proportions are very similar, differences are seen when the proportions of items by area are examined within the test. Further exploration of these differences is required. Some possible reasons for these differences may be due to the artifacts of the simulation procedures used in this set of data. In addition, it could be that the size of the sample pool and the sampling error associated with the simulation might have caused these slight differences within the test.

Because the purpose of this test is to place students based on the total score, these differences are not critical to the functioning of the Elementary Algebra test as a placement tool. The overall proportions are similar across versions. Therefore, comparability in relation to the content specifications of the Elementary Algebra test is realized.

The College Board

BEST COPY AVAILABLE

Conclusions and Next Steps

The purposes of this paper were to (1) provide a research plan for moving an operation CAT program from a network to Internet environment, (2) provide some data on the comparability of the network and Internet versions, and (3) illustrate how new guidelines provide a means of revising a research plan. This report has provided an overview of the types of research projects that should accompany the migration of a CAT program to the Internet. The focus of the research was slightly different than moving a paper-and-pencil assessment onto the computer, but the fundamental issues are very similar. The research plan must include projects in the following areas: (1) comparability, (2) psychometric issues, (3) development projects, (4) validity issues (in this case placement validity issues), and (5) audits or opportunities to review and revise the research plan.

In having presented this plan briefly, initial evidence showing the comparability of the network and Internet versions was provided using data from simulations involving the Elementary Algebra test. Specifically, comparability was examined in the areas of (1) ability estimates, (2) scale scores, (3) performance levels, (4) placement recommendations, and (5) content specifications. There was evidence to support the comparability in making comparable placement decisions between the network and Internet versions.

When the total test was examined, there was good comparability between the network and Internet versions based on content specifications. However, both versions had slight differences in represented the expected content specifications. This may be related to the nature of the Elementary Algebra item pool. Because 92% of the items in the pool represented two of the three content areas, some limitations in adequately representing the three content areas may exist.

In addition, as it has been strongly recommended, the items of any CAT are the foundation for a successful CAT (Stocking, 2000). Therefore, some of the slight variations in the proportions of items represented a content area may be due to a limited pool. Replication of these simulations with the other ACCUPLACER™ placement tests is necessary.

Using the Association of Test Publishers Guidelines for Computer-Based Tests, along with the Joint Standards, a person responsible for the development and maintenance of the computer-based test may develop and/or revise the research plan, especially if one moves into the Internet environment. For example, standard 3.8 encourages that the security of the delivery of computer-base testing should be afforded by the environment. This becomes a difficult

issue if the Internet is utilized especially in an unproctured environment. Therefore, the elimination of requiring some sort of testing center may not be feasible, yet.

The Guidelines for Computer-Based Tests may affect the research plan by suggesting certain studies that might be undertaken. For example, standard 4.2 indicates that if test scores from different modes of administration are interchangeable, there must be a documentation of this equivalency. Thus, studies examining the comparability of scores, as this paper has begun to provide, is called for by this standard.

The next steps involve the following:

(1) Replicate the simulations with the other tests.
(2) Increase the pool size, and replicate the study using all the items.
(3) Examine the effects of a "live", ongoing count of item exposure.
(4) Continue to use the Joint Standards, the Guidelines, other valuable resources, and a final audit of the assessment program to review and revise the research plan.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, D.C.: American Educational Research Association.

Association of Test Publishers. (2000). Provisional draft CBT guidelines. Washington, D.C.: Author.

Bergstrom, B. & Meehan, P. (2000). Computer adaptive technologies. Invited paper presented at the Association of Test Publishers' Computer Based Testing: Applications for the New Millennium, Monterey, CA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Buchanan, T. & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. British Journal of Psychology, 90(1), 125-144.

Coffee, K., Pearce, J., & Nishimura, R. (1999). State of California: Civil service testing moves into cyberspace. Public Personnel Management, 28(2), 283-300.

College Board. (1997). ACCUPLACER™ program overview: Coordinator's Guide. New York: Author.

Gordon, R. J. (1999). Using computer adaptive testing and multiple measures to ensure that students are placed in courses appropriate for their skills. Paper presented at the North American Conference on the learning Paradigm, San Diego, CA.

McNabb, T. (1990). Course placement practices of American postsecondary institutions. ACT Research Report Series 90-10. Iowa City: ACT.

Meijer, R. R. & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. Applied Psychological Measurement, 23(3), 187-194.

National Center for Educational Statistics. (1997). 1993-94 schools and staffing survey: A profile of policies and practices for limited English students:

Screening methods, program support, and teacher training. (NCES Publication No. 97-472). Washington, D.C.: Author.

National Center for Educational Statistics. (1999). The condition of education. (NCES Publication No. 1999-022). Washington, D.C.: Author.

Reckase, M. D. (1989). Adaptive testing: The evolution of a great idea. Educational Measurement: Issues and Practice, 8, 11-15.

Sampson, Jr., J. P. (1998). Using the Internet to enhance test selection, orientation, administration, and scoring. Invited paper presented at the Annual Meeting of the Association for Assessment in Counseling, Indianapolis, IN.

Sireci, S. G., Patelis, T., Rizavi, S., Dillingham, A. M., & Rodriguez, G. (2000). Setting standards on a computerized-adaptive placement examination. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Stocking, M. (2000). Issues and challenges in test planning and design for CBT. Invited paper presentation at the Association of Test Publishers' Conference on Computer-Based Testing: Applications for the New Millennium, Carmel, CA.

Stocking, M. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17, 277-292.

Swanson, L. & Stocking, M. (1993). A model and heuristic for solving very large item selection problems. Applied Psychological Measurement, 17, 151-166.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1995). LOGIST user's guide (Version 7.1). Princeton, NJ: Educational Testing Service.

Appendix A: Characteristics of Elementary Algebra Test

In the elementary algebra test, 12 questions are presented. These are drawn from three categories. The first, operations with integers and rationals, includes computing with integers, computing with negative rationals, ordering, and absolute values; it provides questions appropriate for those with very minimal skills, and accounts for three to four questions for those who are moderately competent in elementary algebra.

The second category, operations with algebraic expressions, also includes several topics appropriate only for the least skilled; evaluating simple formulas and expressions, and adding and subtracting monomials and polynomials. Other topics are administered to students across all levels of performance; these include multiplying and dividing monomials and polynomials, evaluating positive rational roots and exponents, factoring, and simplifying algebraic fractions. Students scoring near the low end of the scale receive six or seven questions drawn from these topics, while those near the top of the scale receive five.

The final category involves the solution of equations, inequalities, and word problems. Most such problems are too difficult to yield useful information for examinees with minimal skills; hence, such students receive only one or two of them. These problems provide the most discriminating questions for examinees who are more competent in elementary algebra. The most able receive as many as six problems, including problems representing each of the following topics: solving linear equations, solving quadratic equations by factoring, translating written phrases into algebraic expressions, solving verbal problems in an algebraic context including geometric reasoning, and graphing. (College Board, 1993).

Appendix B: Summaries of Pool Item Parameters

Table B-1
Distribution of Item Parameters for the Elementary Algebra Test Item Pool

a	Total	% of Total
0.2-0.3	1	0.58%
0.3-0.4	1	0.58%
0.4-0.5	3	1.73%
0.5-0.6	3	1.73%
0.6-0.7	11	6.36%
0.7-0.8	6	3.47%
0.8-0.9	10	5.78%
0.9-1	7	4.05%
1-1.1	9	5.20%
1.1-1.2	20	11.56%
1.2-1.3	19	10.98%
1.3-1.4	22	12.72%
1.4-1.5	14	8.09%
1.5-1.6	9	5.20%
1.6-1.7	9	5.20%
1.7-1.8	5	2.89%
1.8-1.9	8	4.62%
1.9-2	4	2.31%
2-2.1	9	5.20%
2.1-2.2	3	1.73%
Total	173	

b	Total	% of Total
-5--4.5	1	0.58%
-4.5--4	2	1.16%
-4--3.5	1	0.58%
-3.5--3	2	1.16%
-3--2.5	1	0.58%
-2.5--2	3	1.73%
-2--1.5	3	1.73%
-1.5--1	8	4.62%
-1--0.5	18	10.40%
-0.5-0	30	17.34%
0-0.5	42	24.28%
0.5-1	33	19.08%
1-1.5	22	12.72%
1.5-2	6	3.47%
2-2.5	1	0.58%
Total	173	

c	Total	% of Total
0-0.1	32	18.50%
0.1-0.2	69	39.88%
0.2-0.3	55	31.79%
0.3-0.4	12	6.94%
0.4-0.5	5	2.89%
Total	173	

Appendix C: Summary of Performance Level Statements

Test	Score Level	Proficiency Statement
Elementary Algebra	About 108	*Substantial elementary algebra skills.*
	About 76	*Sufficient elementary algebra skills.*
	About 57	*Minimal elementary algebra skills.*
	About 25	*Minimal pre-algebra skills.*

Appendix D: System Functionality Specifications

The following are criteria that will be used to evaluate the Computer Adaptive Test (CAT) module that is submitted for use in The College Board's Internet version of ACCUPLACER. In order to be fair, since local systems connected to the internet may vary, all data related to these criteria will be collected under controlled conditions using a Sun Ultra 5 running Solaris 2.6 with a 270 MHz CPU and 128 MB of RAM. The expectation is that the CAT module will be evaluated through this specified web server using the APACHE Benchmark application.

1. Psychometrics

1.1 The CAT module must be able to utilize information based on a three-parameter model.
1.2 The CAT module must be able to optimize up to 200 content parameters.
1.3 The CAT module must be able to be constrained to handle fixed-length tests.
1.4 The CAT module can accept multiple item display formats.
1.5 The CAT module can be used for tests representing multiple content areas.

2. Computer Technology and Volume

2.1 The CAT module will be expected to handle a minimum of 8,500 transactions as a rate of 400 concurrent transactions per second with zero failures.
2.2 The average latency based on 8,500 transactions should be about 30-40 milliseconds.
2.3 The CAT module must be scalable to allow for 20,000 simultaneous examinees.
2.4 The CAT module must be able to handle these specified volumes without replicating itself causing excessive web-server memory consumption.

3. Computer Technology

3.1 The CAT module must be designed to work and immediately ready for processing in a multi-user web environment, specifically using the Netscape Enterprise Server/Livewire environment.
3.2 There cannot be any disruptions in data transfer between the CAT module and the other components of the web-based Accuplacer system.
3.3 The CAT module should be callable from a Netscape Enterprise Server application.

3.4 The CAT module should have the ability of accepting current state information from Livewire (e.g., closing after receiving a request from Livewire to know that the request has been processed).
3.5 The CAT module should be set-up to allow the application to track multiple concurrent users.